Customer Segmentation Report

1. Objective

   The goal of this study is to segment customers using clustering methods on the given dataset. This dataset combines information from customer profiles and transactions. The goal is to:

   - Identify comparable client groups.
   - Analyze the performance of the clustering techniques K-Means and Hierarchical Clustering.
   - Examine clustering indicators like the Davies-Bouldin Index (DB Index) and the Silhouette Score.
   - To see clusters, use PCA and dendrograms.

2. Data Preparation

   - Customer data includes information such as the customer's ID and location.
   - Transaction data includes measures such as average transaction value, transaction count, and total purchase value.
   - Actions Taken:
     - ♦ Combining customer-specific transaction data.
     - ♦ Use One-Hot Encoding to encode categorical data (such as region).
     - ♦ To ensure that the focus was on feature-based clustering, the two datasets were pooled and the CustomerID column was deleted before the data was prepared for clustering.

3. Methods Used

   The following clustering techniques were applied:

   - ❖ **K-Means Clustering**:
     - ➢ Used the **Elbow Method** to determine the optimal number of clusters.
     - ➢ The selected number of clusters was **5**.
     - ➢ Evaluated using DB Index and Silhouette Score.
   - ❖ **Hierarchical Clustering**:
     - ➢ Used Ward's linkage method.
     - ➢ Visualized the clustering process using a **dendrogram**.
     - ➢ Selected the number of clusters as **5** for comparison.

4. Results

▪ Evaluation Metrics:-

The performance of each clustering algorithm was assessed using the Davies-Bouldin Index and Silhouette Score:

Algorithm	DB_Index	Silhouette Score
K-Means	0.5375340113528055	0.5220966404236095
Hierarchical Clustering	0.535822698369692	0.5266720067937255

5. Visualizations
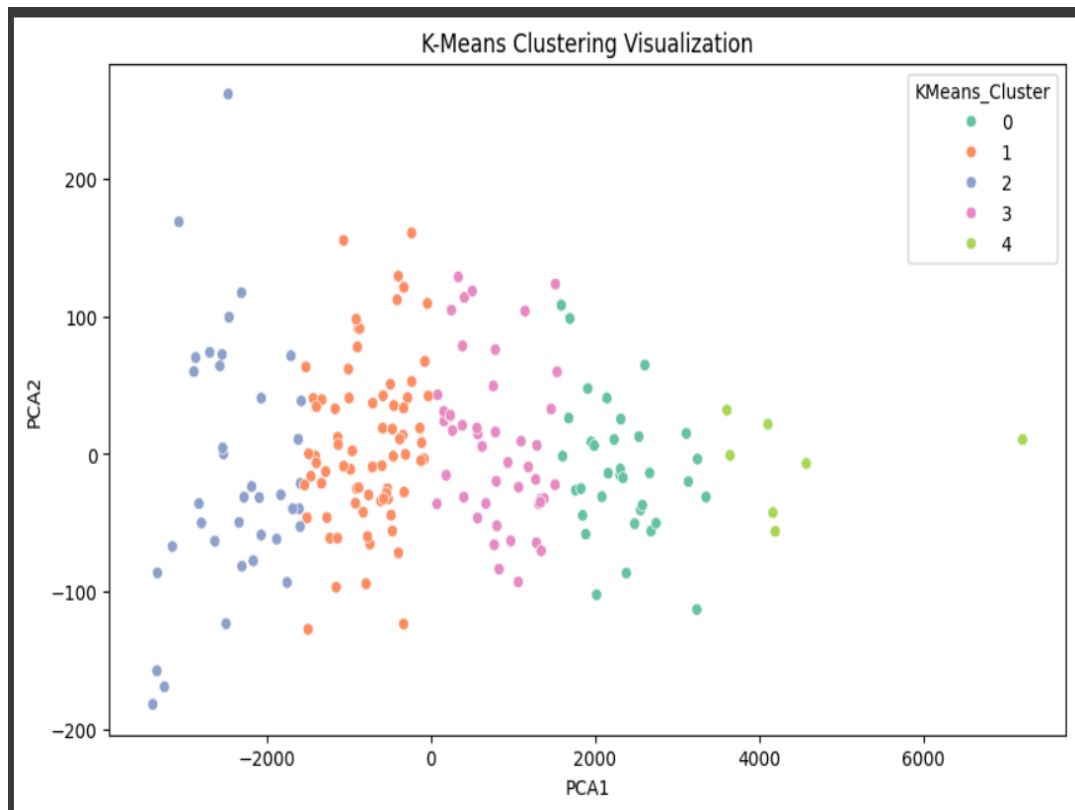
1. **Elbow Method for K-Means**:

   The Elbow Method identified **5 clusters** as the optimal number based on the sharp drop in inertia (SSE).
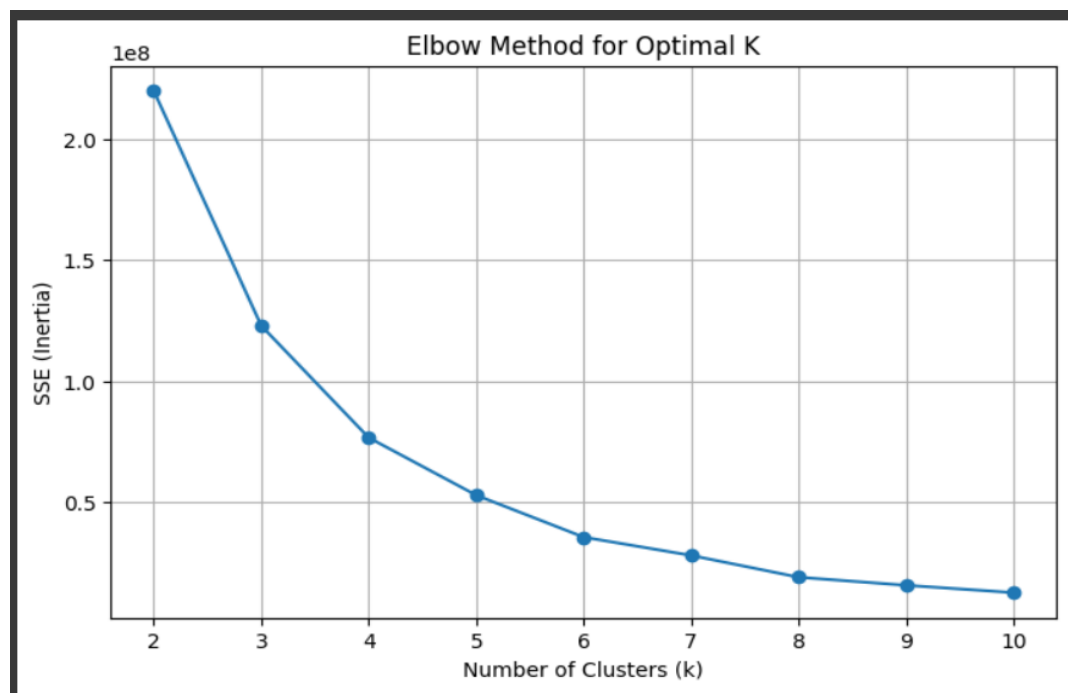
2. **PCA-based Scatterplots**:

   Visualizations of clusters for each algorithm were generated using PCA to reduce dimensions to 2D:

   o **K-Means**: Clear separation among clusters.

K-Means Clustering Visualization

- **Hierarchical Clustering**: Similar separation to K-Means but slightly less distinct.

Hierarchical Clustering Visualization
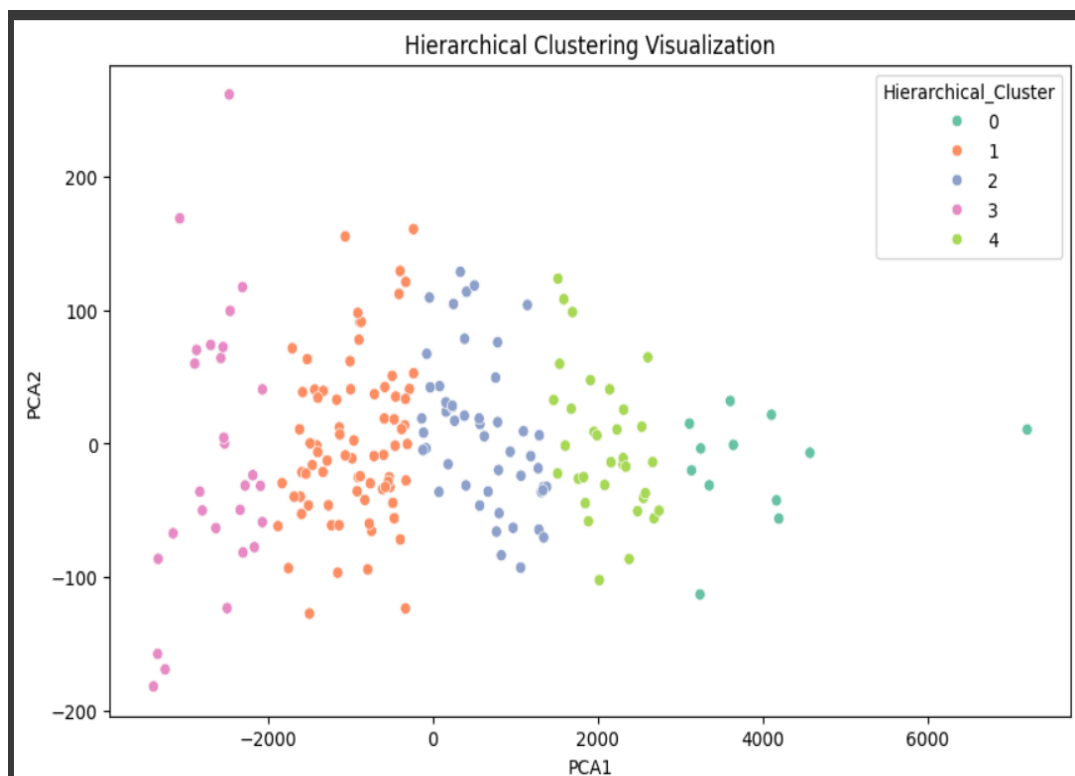
3. **Dendrogram for Hierarchical Clustering**:
   Showed the hierarchical merging of clusters and supported the choice of 5 clusters.